On Composition and Implementation of Sequential Consistency

Matthieu Perrin, Matoula Petrolia, Achour Mostefaoui, Claude Jard

On Composition and Implementation of Sequential Consistency

Matthieu Perrin, Matoula Petrolia, Achour Mostéfaoui, and Claude Jard

LINA – University of Nantes, France first.last@univ-nantes.fr

Abstract. To implement a linearizable shared memory in synchronous message-passing systems it is necessary to wait for a time linear to the uncertainty in the latency of the network for both read and write operations. Waiting only for one of them suffices for sequential consistency.
This paper extends this result to crash-prone asynchronous systems, proposing a distributed algorithm that builds a sequentially consistent shared snapshot memory on top of an asynchronous message-passing system where less than half of the processes may crash. We prove that waiting is needed only when a process invokes a read/snapshot right after a write.
We also show that sequential consistency is composable in some cases commonly encountered: 1) objects that would be linearizable if they were implemented on top of a linearizable memory become sequentially consistent when implemented on top of a sequential memory while remaining composable and 2) in round-based algorithms, where each object is only accessed within one round.

Keywords: Asynchronous message-passing system · Crash-failures · Sequential consistency · Composability · Shared memory · Snapshot

1 Introduction

A distributed system is abstracted as a set of entities (nodes, processes, agents, etc) that communicate through a communication medium. The two most used communication media are communication channels (message-passing system) and shared memory (read/write operations). Programming with shared objects is generally more convenient as it offers a higher level of abstraction to the programmer, therefore facilitates the work of designing distributed applications. A natural question is the level of consistency ensured by shared objects. An intuitive property is that shared objects should behave as if all processes accessed the same physical copy of the object. *Sequential consistency* [1] ensures that all the operations in a distributed history appear as if they were executed sequentially, in an order that respects the sequential order of each process (*process order*).

Unfortunately, sequential consistency is not composable: if a program uses two or more objects, despite each object being sequentially consistent individually, the set of all objects may not be sequentially consistent. *Linearizability* [2] overcomes this limitation by adding constraints on real time: each operation appears at a single point in time, between its start event and its end event. As

a consequence, linearizability enjoys the locality property [2] that ensures its composability. Because of this composability, much more effort has been focused on linearizability than on sequential consistency so far. However, one of our contributions implies that in asynchronous systems where no global clock can be implemented to measure real time, a process cannot distinguish between a linearizable and a sequentially consistent execution, thus the connection to real time seems to be a worthless — though costly — guarantee.

In this paper we focus on message-passing distributed systems. In such systems a shared memory is not a physical object; it has to be built using the underlying message-passing communication network. Several bounds have been found on the cost of sequential consistency and linearizability in synchronous distributed systems, where the transit time for any message is in a range $[d-u, d]$, where d and u are called respectively the *latency* and the *uncertainty* of the network. Let us consider an implementation of a shared memory, and let r (resp. w) be the worst case latency of any read (resp. write) operation. Lipton and Sandberg proved in [3] that, if the algorithm implements a sequentially consistent memory, the inequality $r + w \geq d$ must hold. Attiya and Welch refined this result in [4], proving that each kind of operations could have a 0-latency implementation for sequential consistency (though not both in the same implementation) but that the time duration of both kinds of operations has to be at least linear in u in order to ensure linearizability.

Therefore the following questions arise. Are there applications for which the lack of composability of sequential consistency is not a problem? For these applications, can we expect the same benefits in weaker message-passing models, such as asynchronous failure-prone systems, from using sequentially consistent objects rather than linearizable objects?

To illustrate the contributions of the paper, we also address a higher level operation: a snapshot operation [5] that allows to read in a single operation a whole set of registers. A sequentially consistent snapshot is such that the set of values it returns may be returned by a sequential execution. This operation is very useful as it has been proved [5] that linearizable snapshots can be wait-free implemented from single-writer/multi-reader registers. Indeed, assuming a snapshot operation does not bring any additional power with respect to shared registers. Of course this induces an additional cost: the best known simulation needs $O(n \log n)$ basic read/write operations to implement each of the snapshot operations and the associated update operation [6]. Such an operation brings a programming comfort as it reduces the "noise" introduced by asynchrony and failures [7] and is particularly used in round-based computations [8] we consider for the study of the composability of sequential consistency.

Contributions. We present three major contributions. (1) We identify two contexts that can benefit from the use of sequential consistency: round-based algorithms using a different shared object for each round, and asynchronous shared-memory systems, where programs can not distinguish a sequentially consistent memory from a linearizable one. (2) We propose an implementation of a sequentially consistent memory where waiting is only required when a write is

immediately followed by a read. This extends the result presented in [4] about synchronous failure-free systems, to failure-prone asynchronous systems. (3) The proposed algorithm also implements a sequentially consistent snapshot operation the cost of which compares very favorably with the best existing linearizable implementation to our knowledge (the stacking of the snapshot algorithm of Attiya and Rachman [6] over the ABD simulation of linearizable registers)

Outline. The remainder of this article is organized as follows. In Section 2, we define more formally sequential consistency, and we present special contexts in which it becomes composable. In Section 3, we present our implementation of shared memory and study its complexity. Finally, Section 4 concludes the paper.

2 Sequential Consistency and Composability

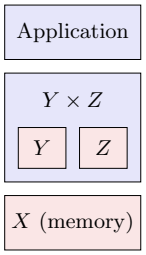
2.1 Definitions

In this section we recall the definitions of the most important notions we discuss in this paper: two consistency criteria, sequential consistency (SC, Def. 2, [1]) and linearizability (L, Def. 3, [2]), as well as composability (Def. 4). A consistency criterion associates a set of admitted *histories* to the *sequential specification* of each given object. A history is a representation of an execution. It contains a set of operations, that are partially ordered according to the sequential order of each process, called *process order*. A sequential specification is a language, i.e. a set of sequential (finite and infinite) words. For a consistency criterion C and a sequential specification T, we say that an algorithm implements a $C(T)$-consistent object if all its executions can be modelled by a history that belongs to $C(T)$, that contains all returned operations and only invoked operations. Note that this implies that if a process crashes during an operation, then the operation will appear in the history as if it was complete or as if it never took place at all.

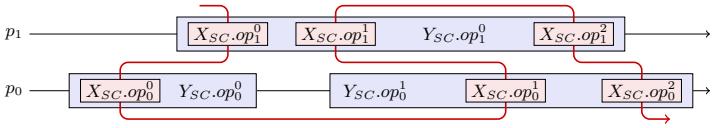
Definition 1 (Linear extension). *Let H be a history and T be a sequential specification. A* linear extension $\leq$ *is a total order on all the operations of H, that contains the process order, and such that each event e has a finite past $\{e' : e' \leq e\}$ according to the total order.*

Definition 2 (Sequential Consistency). *Let H be a history and T be a sequential specification. The history H is* sequentially consistent *regarding T, denoted $H \in SC(T)$, if there exists a linear extension $\leq$ such that the word composed of all the operations of H ordered by $\leq$ belongs to T.*

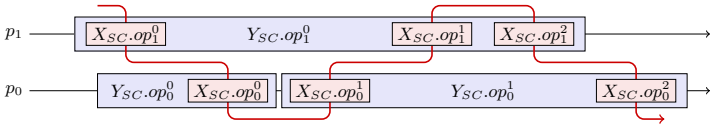
Definition 3 (Linearizability). *Let H be a history and T be a sequential specification. The history H is* linearizable *regarding T, denoted $H \in L(T)$, if there exists a linear extension $\leq$ such that (1) for two operations a and b, if operation a returns before operation b begins, then $a \leq b$ and (2) the word formed of all the operations of H ordered by $\leq$ belongs to T.*



(a) Layer based architecture.



(b) The implementation of upper layer objects call operations on objects from lower layers.



(c) An asynchronous process cannot differentiate this history from the one on Figure 1b.

Fig. 1: In layer based program architecture running on asynchronous systems, local clocks of different processes can be distorted such that it is impossible to differentiate a sequentially consistent execution from a linearizable execution.

Let T_1 and T_2 be two sequential specifications. We define the *composition* of T_1 and T_2, denoted by $T_1 \times T_2$, as the set of all the interleaved sequences of a word from T_1 and a word from T_2. An interleaved sequence of two words l_1 and l_2 is a word composed of the disjoint union of all the letters of l_1 and l_2, that appear in the same order as they appear in l_1 and l_2. For example, the words ab and cd have six interleaved sequences: $abcd$, $acbd$, $acdb$, $cabd$, $cadb$ and $cdab$.

A consistency criterion C is composable (Def. 4) if the composition of a $C(T_1)$-consistent object and a $C(T_2)$-consistent object is a $C(T_1 \times T_2)$-consistent object. Linearizability is composable, and sequential consistency is not.

Definition 4 (Composability). *For a history H and a sequential specification T, let H_T be the sub-history of H containing only the operations belonging to T.*

A consistency criterion C is composable *if, for all sequential specifications T_1 and T_2 and all histories H containing only events on T_1 and T_2, ($H_{T_1} \in C(T_1)$ and $H_{T_2} \in C(T_2)$) imply $H \in C(T_1 \times T_2)$.*

2.2 From Linearizability to Sequential Consistency

Software developers usually abstract the complexity of their system gradually, which results in a layered software architecture: at the top level, an application is built on top of several objects specific to the application, themselves built on top of lower levels. Such an architecture is represented in Fig. 1a. The lowest layer usually consists of one or several objects provided by the system itself, typically a shared memory. The system can ensure sequential consistency globally on all the provided objects, therefore composability is not required for this level. Proposition 1 expresses the fact that, in asynchronous systems, replacing a linearizable object by a sequentially consistent one does not affect the correctness of the programs running on it circumventing the non composability of

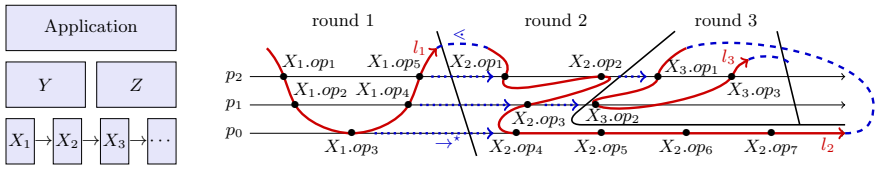
sequential consistency. This result may have an impact on parallel architectures, such as modern multi-core processors and, to a higher extent, high performance supercomputers, for which the communication with a linearizable central shared memory is very costly, and weak memory models such as cache consistency [9] make the writing of programs tough. The idea of the proof is that in any sequentially consistent execution (Fig. 1b), it is possible to associate a local clock to each process such that, if these clocks followed real time, the execution would be linearizable (Fig. 1c). In an asynchronous system, it is impossible for the processes to distinguish between these clocks and real time, so the operations of the objects of the upper layers are not affected by the change of clock. The complete proof of this proposition can be found in [10].

Proposition 1. *Let A be an algorithm that implements an $SC(Y)$-consistent object when it is executed on an asynchronous system providing an $L(X)$-consistent object. Then A also implements an $SC(Y)$-consistent object when it is executed in an asynchronous system providing an $SC(X)$-consistent object.*

An interesting point about Proposition 1 is that it allows sequentially consistent — but not linearizable — objects to be composable. Let A_Y and A_Z be two algorithms that implement $L(Y)$-consistent and $L(Z)$-consistent objects when they are executed on an asynchronous system providing an $L(X)$-consistent object, like on Fig. 1a. As linearizability is stronger than sequential consistency, according to Proposition 1, executing A_Y and A_Z on an asynchronous system providing an $SC(X)$-consistent object would implement sequentially consistent — yet not linearizable — objects. However, in a system providing the linearizable object X, by composability of linearizability, the composition of A_Y and A_Z implements an $L(Y \times Z)$-consistent object. Therefore, by Proposition 1 again, in a system providing the sequentially consistent object X, the composition also implements an $SC(Y \times Z)$-consistent object. In this example, the sequentially consistent versions of Y and Z derive their composability from an anchor to a *common time*, given by the sequentially consistent memory, that can differ from *real time*, required by linearizability.

2.3 Round-Based Computations

Even at a single layer, a program can use several objects that are not composable, but that are used in a fashion so that the non-composability is invisible to the program. Let us illustrate this with round-based algorithms. The synchronous distributed computing model has been extensively studied and well-understood leading the researchers to try to offer the same comfort when dealing with asynchronous systems, hence the introduction of synchronizers [11]. A synchronizer slices a computation into phases during which each process executes three steps: send/write, receive/read and then local computation. This model has been extended to failure prone systems in the round-by-round computing model [8] and to the Heard-Of model [12] among others. Such a model is particularly interesting when the termination of a given program is only eventual. Indeed, some



(a) Round-based program architecture

(b) As the ordering between different objects follows the process order, that is contained into the serialization order of each object, no loop can appear.

Fig. 2: The composition of sequentially consistent objects used in different rounds is sequentially consistent.

problems are undecidable in failure prone purely asynchronous systems. In order to circumvent this impossibility, eventually or partially synchronous systems have been introduced [13]. In such systems the termination may hold only after some finite but unbounded time, and the algorithms are implemented by the means of a series of asynchronous rounds each using its own shared objects.

In the round-based computing model the execution is sliced into a sequence of asynchronous rounds. During each round, a new data structure (usually a single-writer/multi-reader register per process) is created and it is the only shared object used to communicate during the round. At the end of the round, each process destroys its local accessor to the object, so that it can no more access it. Note that the rounds are asynchronous: the processes do not necessarily start and finish their rounds at the same time. Moreover, a process may not terminate a round and keep accessing the same shared object forever or may crash during this round and stop executing. A round-based execution is illustrated in Fig. 2b.

In Proposition 2, we prove that sequentially consistent objects of different rounds behave well together: as the ordering added between the operations of two different objects always follows the round numbering, that is consistent with the program order already contained in the linear extension of each object, the composition of all these objects cannot create loops (Figure 2b). The complete proof of this proposition can be found in [10]. Putting together this result and Proposition 1, all the algorithms that use a round-based computation model can benefit of any improvement on the implementation of an array of single-writer/multi-reader register that sacrifices linearizability for sequential consistency. Note that this remains true whatever is the data structure used during each round. The only constraint is that a sequentially consistent shared data structure can be accessed during a unique round. If each object is sequentially consistent then the whole execution is consistent.

Proposition 2. *Let $(T_r)_{r\in\mathbb{N}}$ be a family of sequential specifications and $(X_r)_{r\in\mathbb{N}}$ be a family of shared objects such that, for all r, X_r is $SC(T_r)$-consistent. Let H be a history that does not contain two operations $X_r.a$ and $X_{r'}.b$ with $r > r'$ such that $X_r.a$ precedes $X_{r'}.b$ in the process order. Then H is sequentially consistent with respect to the composition of all the T_r.*

3 Implementation of a Sequentially Consistent Memory

3.1 Computation Model

The computation system consists of a set Π of n sequential processes, denoted $p_0, p_1, \ldots, p_{n-1}$. The processes are asynchronous, in the sense that they all proceed at their own speed, not upper bounded and unknown to all other processes.

Among these n processes, up to t may crash (halt prematurely) but otherwise execute correctly the algorithm until the moment of their crash. We call a process *faulty* if it crashes, otherwise it is called *correct* or *non-faulty*. In the rest of the paper we will consider the above model restricted to the case $t < \frac{n}{2}$.

The processes communicate with each other by sending and receiving messages through a complete network of bidirectional channels. A process can directly communicate with any other process, including itself (p_i receives its own messages instantaneously), and can identify the sender of the message received. Each process is equipped with two operations: **send** and **receive**.

The communication channels are reliable (no losses, no creation, no duplication, no alteration of messages) and asynchronous (finite time needed for a message to be transmitted but there is no upper bound). We also assume the channels are FIFO: if p_i sends two messages to p_j, p_j will receive them in the order they were sent. As stated in [14], FIFO channels can always be implemented on top of non-FIFO channels. Therefore, this assumption does not bring additional computational power to the model, but it allows us to simplify the writing of the algorithm. Process p_i can also use the macro-operation **FIFO broadcast**, that can be seen as a multi-send that sends a message to all processes, including itself. Hence, if a faulty process crashes during the broadcast operation some processes may receive the message while others may not, otherwise all correct processes will eventually receive the message.

3.2 Single-Writer/Multi-Reader Registers and Snapshot Memory

The shared memory considered in this paper, called a *snapshot memory*, consists of an array of shared registers denoted $\mathsf{REG}[1..n]$. Each entry $\mathsf{REG}[i]$ represents a single-writer/multi-reader (SWMR) register. When process p_i invokes $\mathsf{REG}.\texttt{update}(v)$, the value v is written into the SWMR register $\mathsf{REG}[i]$ associated with process p_i. Differently, any process p_i can read the whole array REG by invoking a single operation namely $\mathsf{REG}.\texttt{snapshot}()$. According to the sequential specification of the snapshot memory, $\mathsf{REG}.\texttt{snapshot}()$ returns an array containing the most recent value written by each process or the initial default value if no value is written on some register. Concurrency is possible between snapshot and writing operations, as soon as the considered consistency criterion, namely linearizability or sequential consistency, is respected. Informally in a sequentially consistent snapshot memory, each snapshot operation must return the last value written by the process that initiated it, and for any pair of snapshot operations, one must return values at least as recent as the other for all registers.

Compared to read and write operations, the snapshot operation is a higher level abstraction introduced in [5] that eases program design without bringing additional power with respect to shared registers. Of course this induces an additional cost: the best known simulation, above SWMR registers proposed in [6], needs $O(n \log n)$ basic read/write operations to implement each of the snapshot and the associated update operations.

Since the seminal paper [15] that proposed the so-called ABD simulation that emulates a linearizable shared memory over a message-passing distributed system, most of the effort has been put on the shared memory model given that a simple stacking allows to translate any shared memory-based result to the message-passing system model. Several implementations of linearizable snapshot have been proposed in the literature some works consider variants of snapshot (e.g. immediate snapshot [16], weak-snapshot [17], one scanner [18]) others consider that special constructions such as test-and-set (T&S) [19] or load-link/store-conditional (LL/SC) [20] are available, the goal being to enhance time and space efficiency. In this paper, we propose the first message-passing sequentially consistent (not linearizable) snapshot memory implementation directly over a message-passing system (and consequently the first sequentially consistent array of SWMR registers), as traditional read and write operations can be immediately deduced from snapshot and update with no additional cost.

3.3 The Proposed Algorithm

Algorithm 1 proposes an implementation of the sequentially consistent snapshot memory data structure presented in Section 3.2. The complete proof of correctness of this algorithm can be found in the technical report [10]. Process p_i can write a value v in its own register $\mathsf{REG}[i]$ by calling the operation $REG.\texttt{update}(v)$ (lines 6-9). It can also call the operation $REG.\texttt{snapshot}()$ (lines 10-11). Roughly speaking, the principle of this algorithm is to maintain, on each process, a local view of the object that reflects a set of *validated* update operations. To do so, when a value is written, all processes label it with their own timestamp. The order in which processes timestamp two different update operations define a *dependency relation* between these operations. For two operations a and b, if b depends on a, then p_i cannot validate b before a.

More precisely, each process p_i maintains five local variables:

- $\mathsf{X}_i \in \mathbb{N}^n$ is the array of most recent validated values written on each register.
- $\mathsf{ValClock}_i \in \mathbb{N}^n$ represents the timestamps associated with the values stored in X_i, labelled by the process that initiated them.
- $\mathsf{SendClock}_i \in \mathbb{N}$ is an integer clock used by p_i to timestamp all the update operations. $\mathsf{SendClock}_i$ is incremented each time a message is sent, which ensures all timestamps from the same process are different.
- $\mathsf{G}_i \subset \mathbb{N}^{3+n}$ encodes the dependencies between update operations that have not been validated yet, as known by p_i. An element $g \in \mathsf{G}_i$, of the form $(g.\mathtt{v}, g.\mathtt{k}, g.\mathtt{t}, g.\mathtt{cl})$, represents the update operation of value $g.\mathtt{v}$ by process $p_{g.\mathtt{k}}$ labelled by process $p_{g.\mathtt{k}}$ with timestamp $g.\mathtt{t}$. For all $0 \leq j < n$, $g.\mathtt{cl}[j]$ contains the timestamp given by p_j if it is known by p_i, and ∞ otherwise.

All updates of a history can be uniquely represented by a pair of integers (k, t), where p_k is the process that invoked it, and t is the timestamp associated to this update by p_k. Considering a history and a process p_i, we define the dependency relation $\rightarrow_i$ on pairs of integers (k, t), by $(k, t) \rightarrow_i (k', t')$ if for all g, g' ever inserted in G_i with $(g.\mathtt{k}, g.\mathtt{t}) = (k, t)$, $(g'.\mathtt{k}, g'.\mathtt{t}) = (k', t')$, we have $|\{j : g'.\mathtt{cl}[j] < g.\mathtt{cl}[j]\}| \leq \frac{n}{2}$ (i.e. the dependency does not exist if p_i knows that a majority of processes have seen the first update before the second). Let $\rightarrow_i^\star$ denote the transitive closure of $\rightarrow_i$.

- $\mathsf{V}_i \in \mathbb{N} \cup \{\perp\}$ is a buffer register used to store a value written while the previous one is not yet validated. This is necessary for validation (see below).

The key of the algorithm is to ensure the inclusion between sets of validated updates on any two processes at any time. Remark that it is not always necessary to order all pairs of update operations to implement a sequentially consistent snapshot memory: for example, two update operations on different registers commute. Therefore, instead of validating both operations on all processes in the exact same order (which requires Consensus), we can validate them at the same time to prevent a snapshot to occur between them. Thus, it is sufficient to ensure that, for all pairs of update operations, there is a dependency agreed by all processes (possibly in both directions). This is expressed by Lemma 1.

Lemma 1. *Let p_i, p_j be two processes and t_i, t_j be two time instants, and let us denote by $\mathsf{ValClock}_i^{t_i}$ (resp. $\mathsf{ValClock}_j^{t_j}$) the value of $\mathsf{ValClock}_i$ (resp. $\mathsf{ValClock}_j$) at time t_i (resp. t_j). We have either, for all k, $\mathsf{ValClock}_i^{t_i}[k] \leq \mathsf{ValClock}_j^{t_j}[k]$ or for all k, $\mathsf{ValClock}_j^{t_j}[k] \leq \mathsf{ValClock}_i^{t_i}[k]$.*

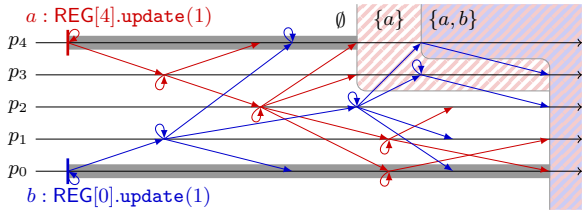
This is done by the mean of messages of the form $\mathtt{message}(\mathsf{v}, \mathsf{k}, \mathsf{t}, \mathsf{cl})$ containing four integers: v the value written, k the identifier of the process that initiated the update, t the timestamp given by p_k and cl the timestamp given by the process that sent this message. Timestamps of successive messages sent by p_i are unique and totally ordered, thanks to variable $\mathsf{SendClock}_i$, that is incremented each time a message is sent by p_i. When process p_i wants to submit a value v for validation, it FIFO-broadcasts a message $\mathtt{message}(\mathsf{v}, i, \mathsf{SendClock}_i, \mathsf{SendClock}_i)$ (lines 8 and 28). When p_i receives a message $\mathtt{message}(\mathsf{v}, \mathsf{k}, \mathsf{t}, \mathsf{cl})$, three cases are possible. If p_i has already validated the corresponding update ($\mathsf{t} > \mathsf{ValClock}_i[\mathsf{k}]$), the message is simply ignored. Otherwise, if it is the first time p_i receives a message concerning this update (G_i does not contain any piece of information concerning it), it FIFO-broadcasts a message with its own timestamp and adds a new entry $g \in \mathsf{G}_i$. Whether it is its first message or not, p_i records the timestamp cl, given by p_j, in $g.\mathtt{cl}[j]$ (lines 14 or 19). Note that we cannot update $g.\mathtt{cl}[\mathsf{k}]$ at this point, as the broadcast is not causal: if p_i did so, it could miss dependencies imposed by the order in which p_k saw concurrent updates. Then, p_i tries to validate update operations: p_i can validate an operation a if it has received messages from a majority of processes, and there is no operation $b \rightarrow_i^\star a$ that cannot be validated. For that, it creates the set G' that initially contains all the operations that have received enough messages, and removes all operations

Algorithm 1: Implementation of a sequentially consistent memory (for p_i)

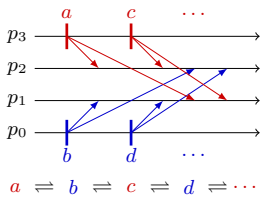
```
   /* Local variable initialization                                          */
 1 X_i ← [0, ..., 0];                          // X_i[j] ∈ ℕ: last validated value written by p_j
 2 ValClock_i ← [0, ..., 0];                   // ValClock_i[j] ∈ ℕ: stamp given by p_j to X_i[j]
 3 SendClock_i ← 0;                                          // used to stamp all the updates
 4 G_i ← ∅;                                  // contains a g = (g.v, g.k, g.t, g.cl) per non-val. update
 5 V_i ← ⊥;                                          // V_i ∈ ℕ ∪ {⊥}: stores postponed updates
   ------------------------------------------------------------------------------------------
   operation update(v) /* v ∈ ℕ: written value; no return value                          */
 6   if ∀g ∈ G_i : g.k ≠ i then                                  // no non-validated update by p_i
 7     SendClock_i++;
 8     FIFO broadcast message(v, i, SendClock_i, SendClock_i);
 9   else V_i ← v;                                                    // postpone the update
   ------------------------------------------------------------------------------------------
   operation snapshot() /* return type: ℕ^n                                               */
10   wait until V_i = ⊥ ∧ ∀g ∈ G_i : g.k ≠ i ;          // make sure p_i's updates are validated
11   return X_i;
   ------------------------------------------------------------------------------------------
   when a message message(v, k, t, cl) is received from p_j
     // v ∈ ℕ: written value, k ∈ ℕ: writer id, t ∈ ℕ: stamp by p_k, cl ∈ ℕ: stamp by p_j
12   if t > ValClock_i[k] then                                       // update not validated yet
13     if ∃g ∈ G_i : g.k = k ∧ g.t = t then                            // update already known
14       g.cl[j] ← cl;
15     else                                                    // first message for this update
16       if k ≠ i then
17         SendClock_i++ ;                                          // forward with own stamp
18         FIFO broadcast message(v, k, t, SendClock_i);
19       var g ← (g.v = v, g.k = k, g.t = t, g.cl = [∞, ..., ∞]); g.cl[j] ← cl;
20       G_i ← G_i ∪ {g};                                // create an entry in G_i for the update
21   var G' = {g ∈ G_i : |{l : g'.cl[l] < ∞}| > n/2};                // G' contains validable updates
22   while ∃g ∈ G_i \ G', g' ∈ G' : |{l : g'.cl[l] < g.cl[l]}| ≠ n/2 do  G' ← G' \ {g'};
23   G_i ← G_i \ G';                                                  // validate updates of G'
24   for g ∈ G' do
25     if ValClock_i[g.k] < g.t then ValClock_i[g.k] = g.t; X_i[g.k] = g.V;
26   if V_i ≠ ⊥ ∧ ∀g ∈ G_i : g.k ≠ i then                          // start validation process for
27     SendClock_i++ ;                                              // postponed update if any
28     FIFO broadcast message(V_i, i, SendClock_i, SendClock_i);
29     V_i ← ⊥;
```

with unvalidatable dependencies from it (lines 21-22), and then updates X_i and $\mathsf{ValClock}_i$ with the most recent validated values (l ines 23-25).

This mechanism is illustrated in Fig. 3a. Processes p_0 and p_4 initially call operation REG.update(1). Messages that have an impact in the algorithm are depicted by arrows and messages that do not appear are received later. The simplest case is process p_3 that received three messages concerning a (from p_4, p_3 and p_2, with $3 > \frac{n}{2}$) before its first message concerning b, allowing it to validate a. The case of p_4 is similar: even if it knows that process p_1 saw b before a, it received messages concerning a from three *other* processes, which allows it to ignore the message from p_1. The situation of p_0 and p_1 may look similar to this of p_4, but the message they received concerning a and one of the messages they received concerning b are from the same process p_2, forcing them to respect the dependency $a \rightarrow_0 b$. The same situation occurs for p_2 so even if a was validated before b by other processes, p_2 must respect the dependency $b \rightarrow_2 a$.



(a) An update is validated by a process when it has received enough messages for this update, and all the other updates it depends of have also been validated.



(b) Infinite chains of dependencies must be avoided to ensure termination.

Fig. 3: Two executions of Algorithm 1

Sequential consistency requires the total order to contain the process order. Therefore, a snapshot of process p_i must return values at least as recent as its last updated value, i.e. it is not allowed to return from a snapshot between an update and the time of its validation (grey zones in Fig. 3a). This can be done in two ways: either by waiting at the end of each update until it is validated, in which case all snapshot operations are done for free, or by waiting at the beginning of all snapshots that immediately follow an update. This extends the remark of [4] to crash-prone asynchronous systems: to implement a sequentially consistent memory it is necessary and sufficient to wait during either read or write operations. In Algorithm 1, we chose to wait during read/snapshot operations (line 10). This is more efficient for two reasons: first, it is not needed to wait between two consecutive updates, which cannot be avoided in the other case, and second the time between the end of an update and the beginning of a snapshot counts in the validation process, but it can be used for local computations. Note that when two snapshot operations are invoked successively, the second one also returns immediately, which improves the result of [4] according to which waiting is necessary for all the operations of one kind.

In order to obtain termination of the snapshot operations (and progress in general), we must ensure that all update operations are eventually validated by all processes. This is expressed by Lemma 2. Figure 3b shows such a case. Process p_2 receives a message concerning a and a message concerning c before a message concerning b, while p_1 receives a message concerning b before messages concerning a and c. This may create dependencies $a \rightarrow_i b \rightarrow_i c \rightarrow_i b \rightarrow_i a$ on a process p_i thus forcing p_i to validate a and c at the same time, even if they are ordered by the process order. Fig. 3b shows that it can result in an infinite chain of dependencies, blocking validation of any update operation. To break this chain, we force process p_3 to wait until a is validated locally before it proposes c to validation by storing the value written by c in a local variable V_i until a is validated (lines 6 and 9). When a is validated, we start the same validation process for c (lines 26-29). Note that, if several updates (say c and e) happen before a is validated, the update of c can be dropped as it will eventually be

overwritten by e. In this case, c will happen just before e in the final linearization required for sequential consistency.

Lemma 2. *If a message* $\texttt{message}(v, i, t, t)$ *is sent by a correct process* p_i*, then beyond some time* t'*, for each correct process* p_j*,* $\mathsf{ValClock}_j^{t'}[i] \geq t$*.*

We can now prove that all histories admitted by Algorithm 1 are sequentially consistent with respect to the snapshot memory object. The idea is to order snapshot operations according to the order given by Lemma 1 on the value of $\mathsf{ValClock}_i$ when they were made and to insert the update operations at the position where $\mathsf{ValClock}_i$ changes because they are validated. This order can be completed into a linear extension, by Lemma 2, and to show that the execution of all the operations in that order respects the sequential specification of the snapshot memory data structure. The complete proof can be found in [10].

3.4 Complexity

In this section, we analyze the complexity of Algorithm 1 in terms of number of messages and latency for each operation. We compare the complexity of our algorithm with the standard implementation of linearizable registers in [15] with unbounded messages. Note that [15] also proposes an implementation with bounded messages but at a much higher cost in terms of latency, which is the parameter we are really interested in improving in this paper. As our algorithm also implements the snapshot operation, we compare it to the implementation of a snapshot object [6] on top of registers. Fig. 4 sums up these complexities.

We measure the complexity as the length of the longest chain of causally related messages to expect before an operation can complete, e.g. if a process sends a message and then waits for some answers, the complexity will be 2.

Each update generates at most n^2 messages and has latency 0, as update operations return immediately. No message is sent for snapshot operations. In terms of latency, in the worst case a snapshot is called directly after two update operations a and b: the process must wait for acknowledgements for its message for a, and then for acknowledgements for its message for b, which gives a complexity of 4. However, if enough time has elapsed between a snapshot and the last update, the snapshot returns immediately.

In comparison, the ABD simulation uses solely a linear number of messages per operation (reads as well as writes), but waiting is necessary for both kinds of operations. Even in the case of the read operation, our worst case corresponds to the latency of the ABD simulation. Moreover, our solution directly implements the snapshot operation. Implementing a snapshot operation on top of a linearizable shared memory is in fact more costly than just reading each register once. The AR implementation [6], that is (to our knowledge) the implementation of the snapshot that uses the least amount of operations on the registers, uses $\mathcal{O}(n \log n)$ operations on registers to complete both a snapshot and an update operation. As each operation on memory requires $\mathcal{O}(n)$ messages and has a latency of $\mathcal{O}(1)$, our approach leads to a better performance in all cases.

	Read		Write		Snapshot		Update	
	# messages	latency	# messages	latency	# messages	latency	# messages	latency
ABD [15]	$\mathcal{O}(n)$	4	$\mathcal{O}(n)$	2	$\sim$	$\sim$	$\sim$	$\sim$
ABD + AR [15,6]	$\sim$	$\sim$	$\sim$	$\sim$	$\mathcal{O}\left(n^2 \log n\right)$	$\mathcal{O}\left(n \log(n)\right)$	$\mathcal{O}\left(n^2 \log n\right)$	$\mathcal{O}\left(n \log(n)\right)$
Algorithm 1	0	$0 - 4$	$\mathcal{O}(n^2)$	0	0	$0 - 4$	$\mathcal{O}(n^2)$	0

Fig. 4: Complexity of several algorithms to implement a shared memory

Algorithm 1, like [15], uses unbounded integer values to timestamp messages. Therefore, the complexity of an operation depends on the number m of operations executed before it, in the linear extension. All messages sent by Algorithm 1 have a size of $\mathcal{O}\left(log(nm)\right)$. The same complexity is necessary to implement n instances of a register with ABD.

In terms of local memory, due to asynchrony, in some cases G_i may contain an entry g for each value previously written. In that case, the space occupied by G_i may grow up to $\mathcal{O}(mn \log m)$. Remark though that, by Lemma 1, an entry g is eventually removed from G_i (in a synchronous system, after 2 time units if $g.\mathtt{k} = i$ or 1 time unit if $g.\mathtt{k} \neq i$). Thus, this maximal bound is unlikely to happen. Also, if all processes stop writing (e.g. in the round based model we discussed in Section 2.3), eventually $\mathsf{G}_i = \emptyset$ and the space occupied by the algorithm drops down to $\mathcal{O}(n \log m)$, which is comparable to ABD. In comparison, the AR implementation keeps a tree containing past values from all registers, in each register which leads to a much higher size of messages and local memory.

4 Conclusion

In this paper, we investigated the advantages of focusing on sequential consistency. We show that in many applications, the lack of composability is not a problem. The first case concerns applications built on a layered architecture and the second example concerns round-based algorithms where processes access to one different sequentially consistent object in each round.

Using sequentially consistent objects instead of their linearizable counterpart can be very profitable in terms of execution time of operations. Whereas waiting is necessary for all operations when implementing linearizable memory, we presented an algorithm in which waiting is only required for read operations when they follow directly a write operation. This extends the result of Attiya and Welch to asynchronous systems with crashes. Moreover, the proposed algorithm implements a sequentially consistent snapshot memory for the same cost.

Exhibiting such an algorithm is not an easy task for two reasons. First, as write operations are wait-free, a process may write before its previous write has been acknowledged by other processes, which leads to "concurrent" write operations by the same process. Second, proving that an implementation is sequentially consistent is more difficult than proving it is linearizable since the condition on real time that must be respected by linearizability highly reduces the number of linear extensions that need to be considered.

Acknowledgments. This work has been partially supported by the Franco-German ANR project DISCMAT under grant agreement ANR-14-CE35-0010-01.

References

1. Lamport, L.: How to make a multiprocessor computer that correctly executes multiprocess programs. Computers, IEEE Transactions on **100**(9) (1979) 690–691
2. Herlihy, M.P., Wing, J.M.: Linearizability: A correctness condition for concurrent objects. ACM Transactions on Programming Languages and Systems (TOPLAS) **12**(3) (1990) 463–492
3. Lipton, R.J., Sandberg, J.S.: PRAM: A scalable shared memory. Princeton University, Department of Computer Science (1988)
4. Attiya, H., Welch, J.L.: Sequential consistency versus linearizability. ACM Transactions on Computer Systems (TOCS) **12**(2) (1994) 91–122
5. Afek, Y., Attiya, H., Dolev, D., Gafni, E., Merritt, M., Shavit, N.: Atomic snapshots of shared memory. J. ACM **40**(4) (1993) 873–890
6. Attiya, H., Rachman, O.: Atomic snapshots in o(n log n) operations. SIAM J. Comput. **27**(2) (1998) 319–340
7. Gafni, E.: Distributed Computing: a Glimmer of a Theory, in Handbook of Computer Science. CRC Press (1998)
8. Gafni, E.: Round-by-round fault detectors: Unifying synchrony and asynchrony (extended abstract). In: Proceedings of the 17$^{\text{th}}$ ACM Symposium on Principles of Distributed Computing, PODC '98, Puerto Vallarta, Mexico. (1998) 143–152
9. Goodman, J.R.: Cache consistency and sequential consistency. University of Wisconsin-Madison, Computer Sciences Department (1991)
10. Perrin, M., Petrolia, M., Mostefaoui, A., Jard, C.: On Composition and Implementation of Sequential Consistency (Extended Version). Research report, LINA-University of Nantes (July 2016)
11. Awerbuch, B.: Complexity of network synchronization. Journal of the ACM **32**(4) (1985) 804–823
12. Charron-Bost, B., Schiper, A.: The heard-of model: computing in distributed systems with benign faults. Distributed Computing **22**(1) (2009) 49–71
13. Dwork, C., Lynch, N., Stockmeyer, L.: Consensus in the presence of partial synchrony. J. ACM **35**(2) (1988) 288–323
14. Birman, K.P., Joseph, T.A.: Reliable communication in the presence of failures. ACM Transactions on Computer Systems (TOCS) **5**(1) (1987) 47–76
15. Attiya, H., Bar-Noy, A., Dolev, D.: Sharing memory robustly in message-passing systems. Journal of the ACM (JACM) **42**(1) (1995) 124–142
16. Borowsky, E., Gafni, E.: Immediate atomic snapshots and fast renaming (extended abstract). In: Proceedings of the Twelth Annual ACM Symposium on Principles of Distributed Computing, Ithaca, New York, USA. (1993) 41–51
17. Dwork, C., Herlihy, M., Plotkin, S.A., Waarts, O.: Time-lapse snapshots. In: Theory of Computing and Systems. Springer (1992) 154–170
18. Kirousis, L.M., Spirakis, P.G., Tsigas, P.: Reading many variables in one atomic operation: Solutions with linear or sublinear complexity. IEEE Trans. Parallel Distrib. Syst. **5**(7) (1994) 688–696
19. Attiya, H., Herlihy, M., Rachman, O.: Atomic snapshots using lattice agreement. Distributed Computing **8**(3) (1995) 121–132
20. Riany, Y., Shavit, N., Touitou, D.: Towards a practical snapshot algorithm. Theor. Comput. Sci. **269**(1-2) (2001) 163–201